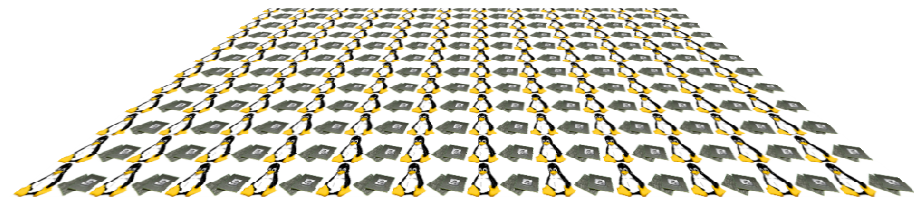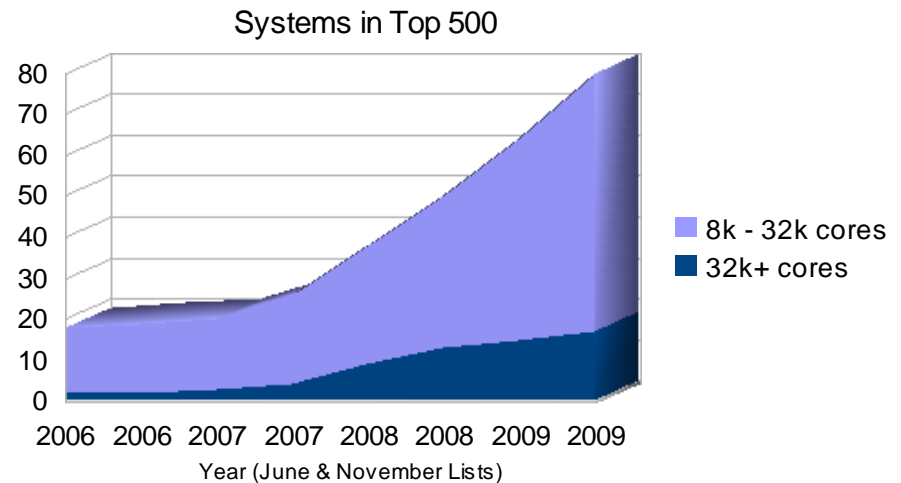**GPU Debugging**

Making it easy

Allinea DDT – CUDA Enabled

- Processor counts growing rapidly

- GPUs entering HPC

- Large hybrid systems imminent

- But what happens when software doesn't work?

Systems in Top 500

8k - 32k cores
32k+ cores

Year (June & November Lists)

- **HPC tools company since 2001**
  - DDT - Debugger for MPI, threaded/OpenMP and scalar
  - OPT - Optimizing and profiling tool for MPI and non-MPI
  - DDTLite – Parallel Debugging Plugin for Microsoft Visual Studio 2008 SP1 and above

- **Large European and US customer base**
  - Ease of use – means tools get used and bugs get fixed - quickly
  - Scalable interface – easy to use at 1 or 100,000s of cores

- **Looking to the future**
  - World's first Petascale debugger
 - GPU product available as pre-release

SCALE TO NEW HEIGHTS

- **Academic**
  - Over 200 universities

- **Major research centres**
  - ANL, EPCC, CEA, IDRIS, Juelich, NERSC, ORNL,

- **Aviation and Defense**
  - Airbus, AWE, Dassault, DLR, EADS, ...

- **Energy**
  - CGG Veritas, IFP, PGS, Total, ..

- **EDA**
  - Cadence, Synopsys, ...

- **Climate and Weather**
  - UK Met Office, IFREMER, Proudman, ...

- **Hybrids are today's hottest topic**
  - Technology is moving quickly – compilers, SDKs, hardware
  - NVIDIA CUDA leads in tool support

- **Many lines of code need rewriting for GPUs**
  - Memory hierarchy
  - Explicit data transfer between host and accelerator
  - Unusual execution model -
    - Kernels, thread blocks, warps, synchronization points
  - Massively fine-grained parallel model

- **Inevitable that we need to debug!**

allinea

- **The first graphical debugger for NVIDIA CUDA**
  - Simple and easy to use
  - As easy as debugging ordinary code

- **Core debugging capability**
  - Breakpoints
  - Stepping warps
  - Viewing data and thread stacks

- **Supports advanced features**
  - CUDA memcheck – memory debugging for CUDA

- **More to come!**

allinea

- **Run the code**
  - Browse source
  - Set breakpoints
  - Stop at a line of CUDA code
  - Stops once for each scheduled collection of blocks

- **Select a CUDA thread**
  - Examine variables and shared memory
  - Step a warp
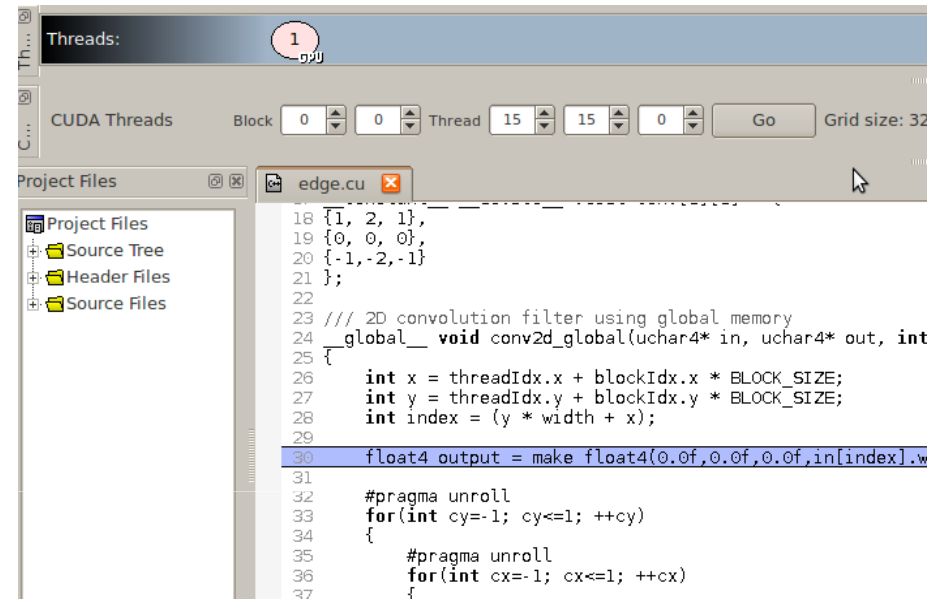
- **View all extant threads in parallel stack view**

  - At one glance, see all GPU and CPU threads together

  - Links with thread selection

  - Pick a tree node to select one of the CUDA threads at that location

- **Full MPI support**

  - See GPU and CPU threads from multiple nodes

allinea

- **Incorrect logic in (if-statements, calculations)**
  - Loop iteration to GPU thread analogy - threads identified by grid and block indexes
  - **Solution:** Select a thread and step with DDT; look at the local state and shared data
    - Cherry-pick important threads: start, end, a few interior points

- **Kernel bounds – getting the right grids and blocks**
  - Incorrect kernel thread boundaries can lead to incomplete results or crashing of the kernel
  - **Solution:** Bugs will often trigger "CUDA memcheck" errors - run with DDT and CUDA memory debugging enabled
  - **Solution:** Use DDT's advanced multi-dimensional array viewer to look at data and find the missing indexes

- **SDK 3.0 is a big leap forward**

- **SDK and driver limitations**

  - Only **one** GPU can be debugged per O/S (per physical node)

  - Cannot currently read launch failure codes (without breaking your code)

  - Only one warp can be stepped per GPU at any time

  - Cannot debug GPU part of (attach to) an already running job

- **Allinea DDT limitations**

  - Not yet possible to compare data across GPU threads

  - "Step threads together" not implemented for CUDA threads

- **A strong partnership with NVIDIA is helping to define future capabilities**